Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 1371 – 1379

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

A Modified Clustering Method Based on Self-Organizing Maps and Its Applications

Le Yang[a], Zhongbin Ouyang[a], Yong Shi[a.b] *

[a]*Research Center on Fictitious Economy & Data Science, the Chinese Academy of Sciences, Beijing 100080,China*
[b]*College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68118, USA*

Abstract

Self-organizing map (SOM) is one of the most popular neural network methods for cluster analysis. Clustering methods using SOM usually are two-stage procedures: first original data are projected onto a set of prototypes on an ordered grid by SOM, and these prototypes can be seen as proto-clusters which will be grouped in the second stage to obtain finally clustering results. Many methods have been proposed to cluster the proto-clusters, among which the prototypes are considered as isolated vectors, without relationship of SOM; others are based on the U-matrix, which represents the local distance structure coming from the topology of SOM. In this paper, we propose a novel method more related to SOM for the purpose to cluster the proto-clusters. In the second stage we use the grid information alternatively, regarding it as a graph partitioned by graph cut algorithm well-known as Normalized cut. We apply this method on image processing and seismic data analysis and obtain reasonable results.

Keywords: Clustering, Self-organizing map, Normalized cut

1. Introduction

Self-organizing map (SOM) has been used for various applications including digital image processing, geosciences and oil exploration, data visualization and others after its introduction in 1982 [1]. The goal of SOM is to represent all input vectors in a high-dimensional space by prototypes in a low-dimensional space, such that the distance and topology are preserved as much as possible.

SOM is one of the most popular neural network methods for cluster analysis [2] for two reasons. First, SOM has the properties of self-organized and topology-preserving. Closed data in the input data space will be projected onto closed prototype vectors on the grid after training phase, so two input vectors which are projected onto closed prototypes are more likely to belong to the same cluster. Second, SOM has prominent visualization properties. The ordered low-dimensional grid can be used as a natural visualization surface to show cluster structure. U-matrix is a common but powerful tool for visualization [3]. After the map is trained, U-matrix will be displayed on top of the grid to give insight into local distance structure between closed prototypes.

* Corresponding author. Tel.: +86 010 82680698; fax: +86 010 82680698.
E-mail address: yshi@gucas.ac.cn.

doi:10.1016/j.procs.2012.04.151

A two-stage procedure of cluster analysis using SOM was proposed in [4]. First the prototypes were produced by SOM. Because of the properties of self-organized and topology-preserving, these prototypes can be seen as "proto-clusters". In the second stage, similar proto-clusters need to be grouped; in fact, clustering of the SOM is needed. Therefore, the key problem is how to categorize the proto-clusters. On this subject, many methods have been proposed, and they can be divided into two types. The first category includes traditional clustering methods such as single linkage and k-means. In these methods, each proto-cluster is simply treated as a single vector, and only these new prototype vectors instead of original input vectors are processed using these methods. The second category is based on the U-matrix, such as flood-fill algorithm on the U-matrix [5], or on U*-matrix (an enhancement of U-matrix) [6] [7].

In this paper, a new method for clustering proto-clusters is proposed. These methods regard the ordered grid as a graph instead and apply graph cut method on it to cluster pro-clusters. The motivation is that during the training phase of SOM, the structure of grid plays a vital role and it should be utilized in the following step. The grid can be seen as a graph, in which every unit is a vertex, and the weight of edge can be calculated by the corresponding prototype vectors, then we will segment the graph to several sub-graphs, namely clusters. This graph-cut problem has been widely studied and there are many well-known methods. In this paper, we choose the Normalized-Cut method [8] and obtain good results.

This paper is organized as follows. In section 2, we review some related work. In section 3, the standard algorithm of SOM is introduced. In section 4, a new two-stage method is proposed. Section 5 gives some experimental results on seismic data analysis and image segmentation. We conclude in Section 6.

2. Articles review

This paper is based on the two-stage procedure frame proposed in [4]. This frame can be introduced as follows: first apply SOM to produce the prototypes, and then cluster the prototypes in the second stage. It is validated to outperform direct clustering methods with less computation time.

Vesanto and Alhoniemi [4] also proposed some methods for clustering the proto-clusters (i.e. single linkage and k-means) and compared their performance. These methods simply take each prototype as a single vector, actually just processing clustering on small scale vectors obtained from SOM. Another way to cluster the proto-clusters is based on the U-matrix. Opolon and Moutarde [5] proposed a semi-automatic method, in which the flood-fill algorithm is applied to segment the U-matrix. Its main idea comes from region growing, a popular method in digital image processing. Because an initial point clearly inside a cluster is needed, this method is semi-automatic. Ultsch [6] proposed a remedy of U-matrix called U*-matrix. Then in [7] the flood-fill algorithm is applied on the U*-matrix, called U*F clustering, which is also semi-automatic. These methods based on U-matrix use the local distance and the structure of SOM, instead of simply setting the prototypes as isolated vectors.

The method proposed in this paper focuses on the regular grid of SOM, regarding it as a graph, and deeply uses the structure of SOM. The problem of clustering turns into a graph-cut problem. For graph cut, an important method based on a minimum cut criterion was proposed in [9]. This method has a drawback that the minimum cut criteria favors cutting small sets of isolated nodes in the graph. Shi and Malik [8] proposed a globe cut criteria based on the minimum criteria called the normalized cut, and used an efficient computational method based on a generalized eigenvalue technique to optimize this criterion. So this paper combines SOM and normalized cut for cluster analysis.

Egmont-Petersena [10] reviewed many applications of neural networks in digital image processing including SOM. Clustering with SOM is also widely used in geosciences and oil exploration [11]. Coléou et.al [12] discussed and compared SOM with other clustering methods in seismic data analysis. In the next section, we will begin with a simply introduction of the standard SOM.

3. SOM training algorithm and U-matrix

The SOM is a regular, usually two-dimensional grid of map units. If d is the dimension of input vector, then each unit is represented by a d-dimensional vector:

$$m_{\mathrm{i}} = \left[m_{i1}, \cdots, m_{id} \right]$$

Because the units are on a regular grid, every one has a neighbourhood relation with others adjacent to it. The neighbourhood relation is decided by the topology of SOM.

The following is SOM's iterative training algorithm. At each step, a sample vector $\boldsymbol{x}$ is randomly selected from the input data set. First compute the distances between $\boldsymbol{x}$ and other prototype vectors. Second, choose the map unit whose corresponding prototype vector is the closest to sample $\boldsymbol{x}$ to be BMU (Best match unit), denoted $\boldsymbol{b}$.

$$b = \arg\min_i \left\{ \left\| x - m_i \right\| \right\}$$

Third, update the prototype vectors. The BMU and its topological neighbours are moved closer to the sample vector in the input space. If the unit is i, then update process will be:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{bi}(t) \left[x - m_i(t) \right]$$

Where t is the number of iterations, $\alpha(t)$ is the learning rate decreasing with t. $h_{bi}(t)$, usually Gaussian function, is descending with the distance between unit i and unit b. Figure 1 is the structure of SOM and Figure 2 presents neighbourhood relationship on the regular hexagon grid, which is usually used.

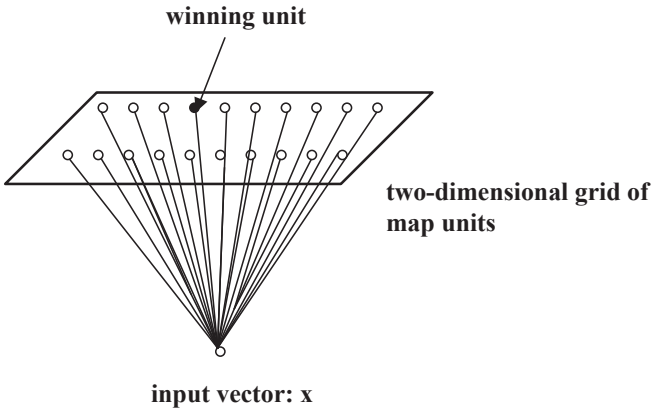


Figure 1: the structure of SOM

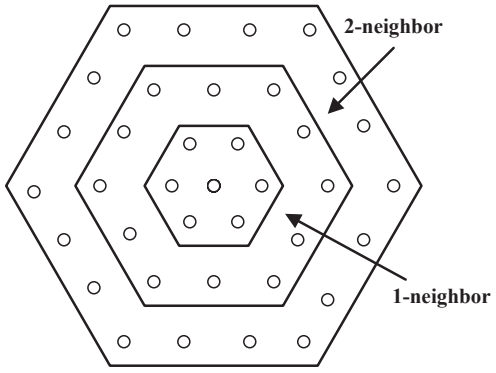


Figure 2: neighbourhood relation on hexagon grid

Two types of SOM are divided by the number of units. If a grid only has a few dozen units, then this SOM's behaviour will look like k-means. In this case, each prototype vector represents a centre of cluster. It is also called k-means SOM. When the number of units is sufficiently large, typically several thousands, the properties of topology

preserving and self-organized play a more vital role. Though large number of units can give more insight to cluster structure of input dataset, the problem is how to visualize it, U-matrix is proposed for this.

A U-matrix (unified distance matrix) is constructed on the trained two-dimensional regular grid. The topological structure of regular grid usually is hexagonal which is convenient to compute distance between near units. If the grid has $n \times m$ units, then U-matrix will be $(2n+1) \times (2m+1)$. Actually, there is a value corresponding to each unit and one of its nearest units, which is the distance between their prototype vectors in the input space, and then the value on the very position of units is the average of its nearest value. Usually, a U-matrix is displayed as a grey level picture, and we can give insight of cluster structure of dataset and choose the cluster manually. The figure of U-matrix will be given in the experiment results. In the next section, we will discuss how to cluster the prototype vectors to gain a final clustering result.

4. Combining SOM and normalized cut for cluster analysis

In the two-stage method proposed in [4], a prototype vector can be interpreted as a "proto-cluster", because closed vectors in the input space will be projected to close BMUs on the grid. Then the proto-clusters will be used to find final clusters in the second stage. As mentioned before, traditional methods such as k-means and single linkage has been used. Since these methods simply take each prototype vector as an isolated point and fail to fully use the structure of SOM, we focus on the methods based on U-matrix in the following.

In the methods based on U-matrix, we can imagine the grey level as height over the grid, so there will be "valleys" where the prototype vectors in the grid are close to each other and "'hills" where the prototype vectors in the grid have a large distance. The U-matrix presents the local distance structure of prototypes trained by SOM and can help us to gain a good clustering result. Simply, we can use the "hills" as boundary to divide the grid to different regions of "valleys" manually. Obviously, selecting clusters manually is a tedious process and nothing guarantees that the manual selection is done consistently. The main idea of semi-automatic algorithm proposed in [5], [6] and [7] is similar. We can use the conception "hills" and "valleys" to demonstrate it. First a start point in a "valleys" as a small region is found, and then the region grows towards every direction, and stop until encountering a "hill". These methods are obviously semi-automatic because of the initial point. Therefore we propose a different way using the ordered grid instead of U-matrix.

From the standard algorithm of SOM, it is obvious that the regular grid plays a vital role in the training step as when the BMU updates, its neighbour units decided by the form of the grid update too. So the grid can also reflect some features of the data, it is reasonable to cluster on proto-clusters with the structure of ordered grid. Clearly, it can be seen as a graph. The vertex is each unit; the edge between two units exists when these two units are adjacent in the ordered grid, and the weight of edges is the distance in the input data space between the prototype vectors on this two units. The cluster is transferred into a graph-partition problem and we choose a well-known method called Normalized cut to solve it.

A simple introduction of Normalized cut (N-cut) methods is as follows. A graph $G = (V, E)$ can be partitioned into two disjoint sets, A, B, $A \cup B = V, A \cap B = \varnothing$ by simply removing edges connecting the two parts. The degree of dissimilarity between these two sets can be computed as total weight of the edges that have been removed. In graph theoretic language, it is called the cut:

$$cut(A, B) = \sum_{u \in A, v \in B} \omega(u, v)$$

A graph cut method is proposed based on minimizing this cut value in [9]. If a graph is to be cut into k subgraphs, recursively using this criterion to bisect the existing segments will be processed. However, this method decline to cut small sets of isolated nodes in the graph. To avoid this unnatural bias for partitioning out small sets of points, a new measure of dissimilarity between two groups called normalized cut is proposed in [8]. The normalized cut criterion is:

$$Ncut(A,B)=\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)}$$

where $assoc(A,V)=\sum_{u\in A,t\in V}\omega(u,t)$. It is a new global criterion for segmenting the graph. The criterion measures both the total dissimilarity between the different groups as well as the total similarity within the groups. An efficient computational technique based on a generalized eigenvalue problem can be used to optimize this criterion and results in practice are very encouraging.

Now we propose our new method combining the SOM and the N-cut for cluster analysis. We use N-cut method to partition the order grid of trained SOM to disjoint sets, and from this partition and the projection of SOM, input vectors can be divided into clusters. N-cut is fit for the task of segmenting the graph of SOM grid; a natural reason is that the grid of SOM also reflects the globe distribution of the whole dataset. On the other hand, the computational load decreases. N-cut method has a drawback that when dataset become very large computational load will be unaccepted, but from the standard algorithm of SOM we can notice that the computational complexity scales linearly with the number of input vectors if the number of units is fixed. In usual applications, the grid with thousands of units is sufficient, so even if the dataset is large, N-cut method is only needed to apply on fixed thousands of proto-clusters, obviously quickly.

The algorithm is described as follows. First the SOM are trained by input vectors and obtain an order grid with pro-clusters. Then the U-matrix can be displayed. Using the information of U-matrix, we can decide the number of disjoint sets the grid will be partitioned to. The N-cut method is used to partition the graph established by both grid and pro-clusters. This step gets a cluster result of the proto-clusters. Finally, by inversing the projection of SOM, we can get the cluster results of the input vectors. The whole process is, obviously, automatic. In the next section, some experiment results are shown to demonstrate this method.

5. Experimental results

5.1. Image segmentation

To exemplify the effectiveness of the proposed algorithm, we use it on the task of image segmentation based on pixel data. An image can be seen as a matrix I, where $I(x,y)$ is the gray value in row x and column y of the image. Our samples for training are three-dimensional vectors, comprised of spatial coordinates and gray values, namely (x,y,z), z represents the gray value of pixel (x,y).

First we present a simple logic image which only has two pixel value of {0, 1}. Figure 3a is the original picture. Figure 3c and 3d is the U-matrix and regular grid of trained SOM, obviously it can be divided into two parts, so we can constrain the number of subgraphs is two. Figure 3b is the result of whole process. Because of the grid is well-separated, the result is perfect.

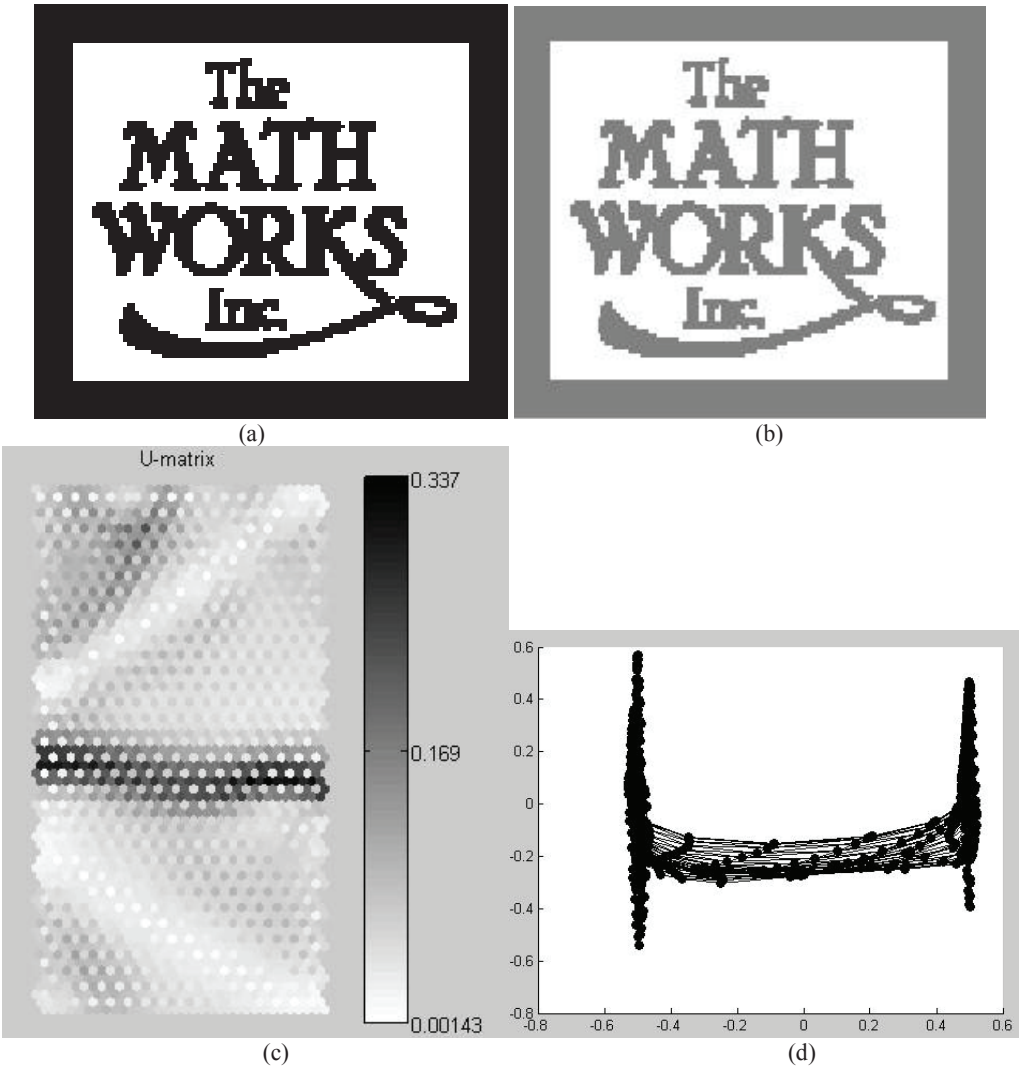


Figure 3: "The Math Works"

Then we apply it on a more complex gray image. Note that because of the ability of SOM, for some images, the U-matrix after training is not well-separated. In this case, our method can give a more sensible result while it is a very difficult task to choose cluster manually. Figure 4a is the original picture, Figure 4b is the segmentation result Figure 4c and 4d denote the U-matrix and regular grid of trained SOM, and we can see all of them is complicated and can not choose cluster reasonably. So we use the N-cut method to divide it into many parts and get the result displayed in Figure 4b. From the result we can see although the U-matrix and grid are complicated, N-cut method can obtain a segmentation reflecting the clustering structure of original picture.

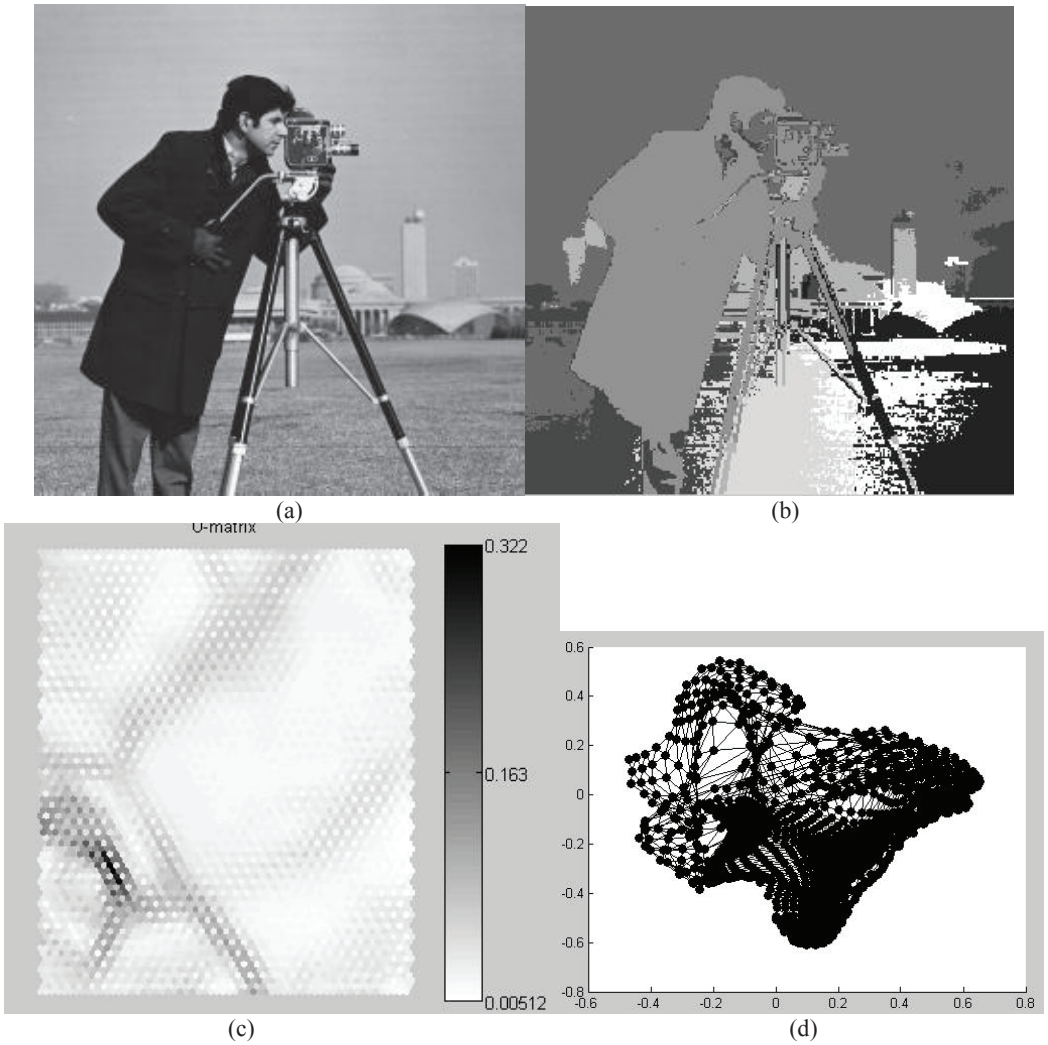


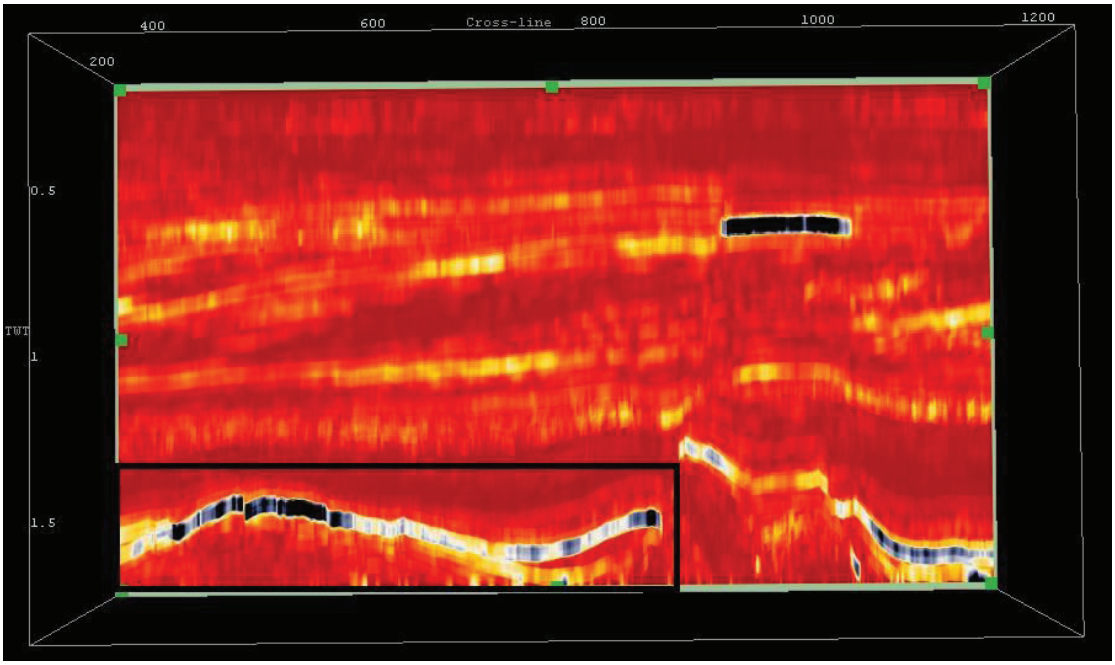
Figure 4: "The Cameraman"

5.2. Seismic data

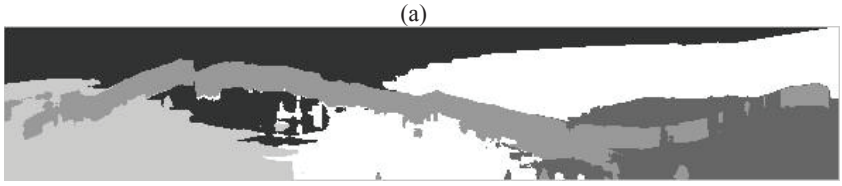
In seismic data processing, one of the goals is to find some region where the attributes are alike. For example, given original seismic data, we can compute the value of energy on every position, then if we can find a region in which energy are notable higher, it will be called "hot spot". Natural gas around the hot spot then may be found, so our results can give a reference to experts in oil exploration. Note that this process is similar to image segmentation, every position has a spatial coordinate and an attribute value, and we also need segmentation. These results can be further used by optimization based methods to obtain more complex and useful information which can help us work better [13].

The data we use come from famous free software named OpendTect, this software provide a demo data, this demo has been analyzed by experts and give complete results, and we apply our method on it for validation.

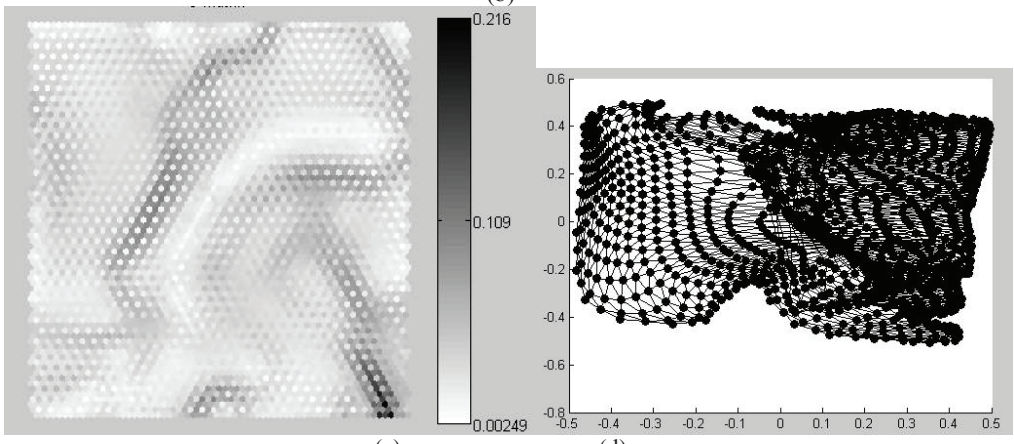
Figure 5a is the original seismic data. For simplification, we choose a small region of original data in the lower left. Because the data demo has been analyzed, we can see this region comprise a sigmoid hot spot structure. Figure 5c and 5d is the U-matrix and regular grid of trained SOM, also confusing. Figure 5b is the final result; it is easy to see a sigmoid structure. But other region with similar attribute is divided into several parts; the reason we use the Normalized cut is that this method inclines to divide the graph to subgraphs which have almost similar scale.



(a)



(b)



(c) (d)

Figure 5: the seismic data

6. Conclusion

In this paper we proposed a clustering method combining the SOM and normalized cut algorithm. The regular grid of trained SOM was considered as a graph and use normalized-cut criterion to segment it, each subgraph was a cluster. From the experiments we can see that when the U-matrix is well-separated, our algorithm can get as good clustering results as that based on segmenting the U-matrix. However, if the U-matrix is complicated, segmentation will be infeasible because the initial point can not be found consistently, while our method still can gain reasonable

results. In the future work, more graph-cut algorithms will be combined with SOM for cluster analysis.

Acknowledgements

This research has been partially supported by grants from BHP Billiton Co., Australia, CAS/SAFEA International Partnership Program for Creative Research Teams (#70921061), and National Natural Science Foundation (#90718042).

References

1. T. Kohonen (1982), Self-Organized formation of topologically correct feature maps, Biological Cybernetics, vol. 43, p.59-69.

2. Jiawei Han, Micheline Kamber and Jian Pei (2011), Data Mining: Concepts and Techniques, Third Edition, Mogan Kaufmann.

3. A. Ultsch & H.P. Siemon (1990), Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis, In Proc. Intern. Neural Networks Conf. (INNC '90), Dortrecht (Netherlands), Kluwer Academic Press, Paris, p. 305-308.

4. J. Vesanto and E. Alhoniemi (2000), Clustering of the Self-Organizing Map, IEEE Transactions on Neural Networks, vol. 11 (3).

5. D. Opolon & F. Moutarde (2004), Fast semi-automatic segmentation algorithm for Self-Organizing Maps, In Proc. of ESANN'2004, Bruges, 28-30 avril 2004, p. 507-512.

6. A. Ultsch (2003), U*-Matrix: a Tool to visualize Clusters in high dimensional Data, In Research report Dept. of Mathematics and Computer Science, University of Marburg(Germany), No. 36.

7. Fabien Moutarde and Alfred Ultsch (2005), U*F Clustering: A New Performant "Cluster-Ming" Method Based on Segmentation of Self-Organizing Maps, Workshop on Self-Organizing Maps (WSOM'2005), Paris : France.

8. Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 888–905.

9. Z. Wu and R. Leahy (1993), An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1,101-1,113, Nov. 1993.

10. M. Egmont-Petersena , D. de Ridder and H. Handels (2002), Image processing with neural networks—a review,Pattern Recognition 35 (2002) 2279–2301.

11. Masoud Nikravesh (2007), Computational Intelligence for Geosciences and Oil Exploration,Studies in Fuzziness and Soft Computing, 2007, Volume 217/2007, 267-332, DOI: 10.1007/978-3-540-73182-5_14.

12. Thierry Coléou, Manuel Poupon and Kostia Azbel (2003), Unsupervised seismic facies classification: A review and comparison of techniques and implementation, The Leading Edge, October 2003, v. 22;,no. 10,p. 942-953, DOI: 10.1190/1.1623635.

13. Yong Shi, Yingjie Tian, Gang Kou, Yi Peng, Jianping Li (2011),Optimization Based Data Mining: Theory and Applications, Springer